INTRODUCTION

This report contains a number of chapters, each of which can be read independently. For a number of practical reasons we decided to present the report in this form, rather than to impose an artificial continuity by substantial rewriting. Dr David Harper, formerly a research student of van Rijsbergen, was a co-author of Chapter 2, which is to be published in Information Processing and Management. Chapter 6 will be published in Program, and Chapter 3 in the Proceedings of the BCS-ACM Conference on Research and Development in Information Retrieval held in Cambridge in June, 1980.

The aim of the project, broadly speaking, was to investigate methods of predicting which terms would be useful in searching and how useful they would be. To some extent both van Rijsbergen and Robertson had been working independently on this for several years now, developing and testing probabilistic models, and using samples obtained by relevance feedback to estimate the usefulness of certain index terms. For some time it was thought that

1) the tests were made on inadequate data, and

2) the relationship between the Rijsbergen/Robertson probabilistic approach and the Harter 2-Poisson model should be investigated.

The hope was that (2) might lead to a more general probabilistic model, one which would encompass the 2-Poisson distribution model for the frequency of occurrence of terms within documents.

To meet (1), we established at Cambridge a test collection of appropriate size. The original data for this came from the National Physical Laboratory where, in the sixties, Vaswani and Cameron had created a machine readable set of abstracts, queries and relevance assessments. The original data and some of the problems it presented in processing are described in Chapter 1. There was already a program suite at Cambridge for setting up test collections, but it was decided to rewrite it, if only for the purpose of creating the new NPL test collection. A description of these programs, together with their use in setting up the NPL test collection, is given in Chapter 5. In particular, a new suffix stripping program was designed and implemented as part of the package. This is fully described in Chapter 6.

The NPL test collection, derived from abstracts, was somewhat different from previous test collections in that each term had associated with it a frequency count representing the number of times that term occurred in the abstract. This frequency information was needed to investigate the Harter model and its relationship with our probabilistic work. Chapter 3 contains a brief description (and introduction) to Harter's work. One should also look there for further references. Harter had proposed a 2-Poisson distribution to model the statistical distribution of content bearing words. His original aim was to use these for selecting index terms. Our aim was to investigate whether his approach could be incorporated into the Rijsbergen/Robertson probabilistic approach and to test whether this would lead to improvement in retrieval effectiveness. This work constituted a generalisation of our previous work, since in the past we had been mainly concerned with binary data, and we wanted to extend our work to frequency data. The results of this 'marriage' and its empirical testing can be found in Chapter 3. Some further theoretical lines of work which, for various reasons, did not reach any kind of conclusion are discussed in Chapter 4.

As a separate development, Rijsbergen pursued his original work on the usefulness of co-occurrence data. This work required the implementation of an efficient spanning tree algorithm able to cope with the large data sets. Porter redesigned an existing algorithm for this purpose and implemented it, and a description can be found in the Appendix to Chapter 2. With the aid of this fast spanning tree algorithm, a large number of experiments were conducted to investigate the effectiveness of query expansion when the extra search terms are derived from a spanning tree based on a number of association measures. Most of the results of this work are reported in Chapter 2. Some negative results are reported in Chapter 3. Some further theoretical lines of work which, for various reasons, did not reach any kind of conclusion, are described in Chapter 4.

In the course of the project it soon became apparent that to conduct our experiments in an organised manner and on the scale required, some high level software would be needed. In other words we needed an experimental IR system able to cope conveniently with the large number of experiments that we needed to run. Chapter 7 gives a description of what is by now a subset of the final system written by Porter for this purpose. The software described in this report (with the possible exception of the spanning tree algorithm) is such that it would not be too difficult to implement it on some other installation.

Acknowledgements

The work described in this report was carried out under a grant from British Library Research and Development (grant no. SI/G/300) to Professor M.V. Wilkes, Computer Laboratory, University of Cambridge. We must thank the University Computing Service for the generous supply of computing resources over the two year period 1979-1980. Van Rijsbergen was supported by a Royal Society Research Fellowship throughout the project.

The main collection of data used in the experiments was supplied through Dr Peter Vaswani of the National Physical Laboratory, Teddington, and we are greatly obliged to him for making this material available.

We must also thank Dr David Harper and Dr Karen Sparck Jones for their many

suggestions and criticisms throughout the course of this project.

Keith van Rijsbergen
Stephen Robertson
Martin Porter

September 1980